

Potential Theory for Directed Networks

Qian-Ming Zhang¹, Linyuan Lü^{2,*}, Wen-Qiang Wang¹, Yu-Xiao Zhu¹, and Tao Zhou¹

¹*Web Science Center, School of Computer Science and Engineering,
University of Electronic Science and Technology of China,
Chengdu 611731, People's Republic of China*

²*Department of Physics,
University of Fribourg, Chemin du Musée 3
Fribourg CH-1700, Switzerland*

(Dated: February 14, 2012)

Uncovering mechanisms underlying the network formation is a long-standing challenge for data mining and network analysis. In particular, the microscopic organizing principles of directed networks are less understood than those of undirected networks. This article proposes a hypothesis named *potential theory*, which assumes that every directed link corresponds to a decrease of a unit potential and subgraphs with definable potential values on all nodes are preferred. Combining the potential theory with the clustering and homophily mechanisms, we deduce that the Bi-fan consisting of 4 nodes and 4 directed links is the most favoured local structure in directed networks. Our hypothesis get positive supports from extensive experiments on 15 directed networks drawn from disparate fields, as indicated by the most accurate and robust performance of Bi-fan predictor within the link prediction framework. In summary, our main contributions are twofold: (i) We propose a new mechanism for the organization of directed networks; (ii) We design the corresponding link prediction algorithm, which can not only testify our hypothesis, but also find direct applications in missing link prediction and friendship recommendation.

PACS numbers: 89.75.Fb, 05.40.Fb, 89.75.Da

I. INTRODUCTION

Many social, biological and technological systems can be well described by networks, where nodes represent individuals and links denote the relations or interactions between nodes. The study of structure and functions of networks has therefore become a common focus of many branches of science [1]. A big challenge attracting increasing attention in the recent decade is to uncover the mechanisms underlying the formation of networks [2]. Macroscopic mechanisms include the rich-get-richer [3], the good-get-richer [4], the stability constrains [5], and so on. Microscopic mechanisms include homophily [6], clustering [7], balance theory [8], and so on. Mechanisms can also play a part in regulating the mesoscopic structure, like the formation and transformation of groups and communities [9–11]. Real networks are usually resulted from a hybrid of several mechanisms, for example, new nodes may form links according to the rich-get-richer mechanism, and simultaneously, new links among old nodes could be a consequence of the mechanism of clustering [12].

A number of systems are naturally described by directed networks: the world wide web is made up of directed hyperlinks, the food webs consist of directed links from predators to preys, and in the microblogging social networks, fans form links pointing to their opinion leaders. The formation of directed links also obey some

general mechanisms, for example, users in Twitter are likely to form links to neighbors of their neighbors and to users with near ages, which are in accordance with the clustering and homophily mechanisms [13]. Reciprocity is a specific mechanism for directed networks [14]. It is valid for many social networks, but inapplicable for some others like food webs. Compared with undirected networks, link formation in directed networks receives less attention and thus has not been well understood.

In this article, we propose a hypothesis on link formation for general directed networks, named *potential theory*. Combining the potential theory and the clustering and homophily mechanisms, we could deduce a certain preferred subgraph. We apply the link prediction approach [15] to verify our deduction. That is, we hide a few fraction of links and predict them by assuming that a link generating more preferred subgraphs is of higher probability to exist. Experiments on disparate directed networks ranging from large-scale social networks containing millions of individuals to small-scale food webs consisting of a hundred of species show that the prediction according to the preferred subgraph is remarkably more accurate and robust than prediction according to other comparable subgraphs. Besides the insights of the underlying mechanism for directed network formation, our work could find applications in friendship recommendation for social networks and missing link prediction for biological networks.

This article is organized as follows. In Section 2, we will introduce some closely related works. Our perspectives and methods are presented in Section 3. The data description, experimental results and analyses are shown

*Electronic address: linyuan.lue@unifr.ch

in Section 4. Lastly, in Section 5, we summary the main finding and outline implications of this work.

II. RELATED WORK

As we will show in Section 3, the potential theory is locally applicable and thus we only introduce the previous work on microscopic mechanism. In addition, we also introduce the studies on link prediction, emphasizing the applications of link prediction approaches on mechanism evaluation.

A. Microscopic Mechanisms

Clustering mechanism declares that two nodes have high probability to make a link between them if they share some common neighbors. This mechanism is indirectly supported by increasing evidences on high clustering coefficients (the clustering coefficient of a node is defined as the density of links among its neighbors, and the clustering coefficient of the whole network is the average over all nodes [16]) of disparate networks [7]. Through investigation on a social network comprising 43,553 university members, Kossinets and Watts [17] found direct evidence that two students share more common acquaintances are more likely to become acquaintance to each other. The clustering mechanism also works for directed networks, for example, in Twitter, more than 90% of new links are added between nodes having at least one common neighbor [13]. In addition, evolving network models driven by common neighbors could reproduce some significant features of both real directed and undirected networks [18, 19].

Homophily mechanism states the observed tendency of people to associate with others of similar profiles and/or experiences [6]. Experiments on social networks thus far strongly support this mechanism. Positive evidences come from various channels, such as an acquaintance network of university members [17], a large-scale instant-messaging network containing 1.8×10^8 individuals [20], friendship networks of a set of American high schools [21], a social network of a cohort of college students in Facebook [22], and so on. A variety of characteristics are shown to be significant to the link formation, including race, tastes in music and movies, grade, age, location, language, sharing experience, etc. Homophily mechanism also plays a role in other kinds of networks, for example, in directed document networks, links (e.g., hyperlinks between web pages and citations between articles) tend to connect documents with high content similarities [23]. In some literature, the clustering mechanism is considered as a special case of homophily mechanism, where two nodes having some common neighbors are recognized as being of similar network surroundings. We prefer to distinguish these two mechanisms. Recent experiments on directed social networks show that the

clustering mechanism may be even stronger than the homophily mechanism [24].

Reciprocity mechanism is the tendency of nodes to response to incoming links by creating links back to the source [14]. It is a specific mechanism for directed networks, but not applicable everywhere. For example, the reciprocity mechanism plays a significant role in the growth of social networks of Facebook [25] and Flickr [26], while it is of much less impacts on Slashdot [27] and it does not work at all for food webs [28].

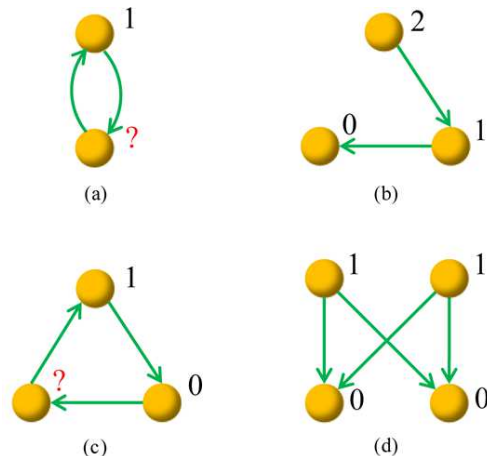


FIG. 1: Illustration of four example graphs. Graphs (b) and (d) are potential-definable, and the numbers labeled beside nodes are example potentials. Graphs (a) and (c) are not potential-definable, and if we set the top nodes' potential to be 1, some nodes' potentials cannot be determined according to the constrain that a directed link is always associated with a decrease of a unit potential.

B. Link Prediction

Link prediction is one of the most fundamental problems in networked data mining that attempts to estimate the likelihood of the existence of a link between two nodes, according to observed links and the attributes of nodes [15, 29]. It has found applications in predicting missing links of biological networks and recommending social mates in online social networks. To evaluate the algorithmic performance, we usually divide the data into two parts and use the training set to predict the testing set. The algorithmic accuracy is quantified by counting the overlap of prediction and the testing set and/or the ranks of links in the testing set among all non-observed links (i.e., node pairs not in the training set).

Link prediction algorithm can be used in judging the driven mechanisms of network formation. The very nice performance of common-neighbor-based similarity indices strongly supports the validity of clustering mechanism [30, 31]. On several real friendship networks, Aiello et al. [32] showed that when combined with topological

features, topical similarity achieves a link prediction accuracy of about 92%. Via a link prediction approach, Wang et al. [33] found that the mobility homophily and structural clustering are both significant for link formation. Leskovec et al. [34] testified the validity of social balance theory through link prediction in signed networks.

III. METHODS

This section will present our methods in details. Firstly, we will introduce the perspectives of potential theory, and then deduce the preferred subgraph by combining the potential theory together with clustering and homophily mechanisms. Lastly, we will build a bridge between the preferred subgraph and the link prediction problem, say how to testify our hypothesis by using a link prediction algorithm.

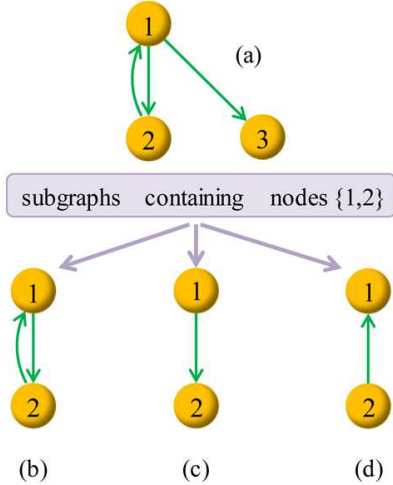


FIG. 2: Considering subgraphs of (a) that contains nodes $\{1,2\}$, according to the traditional definition, (b) is the unique one, while in our definition, graphs (b), (c) and (d) are all subgraphs. Notice that, the empty graph containing nodes 1 and 2 and no link is also a subgraph of (a) according to our definition.

A. The Potential Theory

A graph is called potential-definable if each node can be assigned a potential such that for every pair of nodes i and j , if there is a link from i to j , then i 's potential is a unit higher than j . Clearly, a link is potential-definable yet a graph containing reciprocal links is not potential-definable. Figure 1 illustrates some example graphs with orders from 2 to 4, where graphs (a) and (c) are not potential-definable and graphs (b) and (d) are potential-definable.

The potential theory claims that a link that can generate more potential-definable subgraphs is more significant and thus of higher probability to appear. Our definition of subgraph is more general than traditional one. Given a directed graph $\mathbb{D}(V, E)$ with V and E the sets of nodes and directed links, according to the traditional definition, a graph $\mathbb{D}'(V', E')$ is called a subgraph of \mathbb{D} if $V' \subset V$ and E' contains all the links in E that connect two nodes in V' . Our definition only requires $V' \subset V$ and $E' \subset E$, that is, E' is not necessary to include all links connecting nodes in V' . As shown in figure 2, (b), (c) and (d) are subgraphs of (a) according to our definition, yet only (b) is a subgraph of (a) in the traditional definition.

B. The Preferred Subgraph

Since any graph containing reciprocal links is not potential-definable, here we do not take into account the reciprocity mechanism. The clustering mechanism prefers short loops (not necessary to be directed loops) and it only works for local surrounding, and thus we only consider loop-embedded subgraphs with orders 3 and 4. Two nodes connected by reciprocal links are not treated as loops. To avoid the repeated count, we only consider the minimal loop-embedded subgraphs that do not contain loop-embedded subgraphs themselves.

Figure 3 illustrates all the six different minimal loop-embedded subgraphs of orders 3 and 4. These subgraphs are named after Ref. [35] yet our motivation is different from motif analysis and we adopt a different definition about subgraph. Among these six subgraphs, only Bi-fan and Bi-parallel are potential-definable. In a potential-definable subgraph, two nodes with the same potential cannot directly connect to each other and thus the homophily mechanism only works when we consider the subgraphs as a whole. For Bi-fan the links are equivalent to each other and nodes are of two different potentials, while in Bi-parallel, links are different (two are from high-potential nodes to moderate-potential nodes, and the other two are from moderate-potential nodes to low-potential nodes) and nodes are of three different potentials. According to the assigned potentials, we could say the Bi-fan structure is more homogeneous than the Bi-parallel structure, and thus the homophily mechanism prefers the former one.

In a word, taking into account the potential theory, together with the clustering and homophily mechanisms, we think that the Bi-fan subgraph is the most preferred one and thus a link that can generate more Bi-fan structure should be of higher probability to exist.

C. Link Prediction Algorithm

Given a directed network $\mathbb{D}(V, E)$, the fundamental task of a link prediction algorithm is to give a rank of

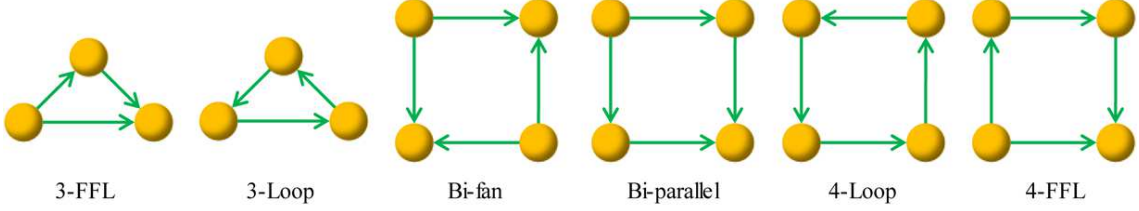


FIG. 3: All the six minimal loop-embedded subgraphs of orders 3 and 4. They are named after Ref. [35], where 3-FFL and 4-FFL stand for three-order and four-order feed forward loops, and 3-Loop and 4-Loop mean three-order and four-order feedback loops, respectively.

all non-observed links in the set $U \setminus E$, where U is the universal set containing all $|V|(|V| - 1)$ possible directed links. If one wants to find out missing links or recommend friendships, one can go for the links with the highest ranks. The mainstream method is to assign each non-observed link a score, and the one with higher score is ranked higher.

We design the predictors corresponding to the six minimal loop-embedded subgraphs shown in figure 3. By removing one link from every subgraph, we get twelve predictors as shown in figure 4. If we adopt the predictor S_i , it means the score of a non-observed link $u \rightarrow v$ is defined as the number of the i th subgraphs created by the addition of this link. Notice that, a link may generate ten 3-FFLs, but its roles can be different. For example, these ten 3-FFLs may include two S_1 , three S_2 and five S_3 . So if we adopt the predictor S_2 , the score of this link is three. Therefore, if we would like to see the contribution of a link to the created 3-FFLs, we can adopt the predictor $S_1 + S_2 + S_3$, which means that the score of a non-observed link is defined as the sum of created S_1 , S_2 and S_3 by this link, equivalent to the number of created 3-FFLs. Figure 5 illustrates a simple example about how we calculate the scores.

Given a predictor we can rank all the non-observed links according to their scores. To evaluate the algorithmic performance, we randomly divide the observed links E into two parts: the training set E^T is treated as known information, while the testing set (probe set) E^P is used for testing and no information therein is allowed to be used for prediction. Clearly, $E = E^T \cup E^P$ and $E^T \cap E^P = \phi$. In our experiments, the training set always contains 90% of links, and the remaining 10% of links constitute the testing set.

We use a standard metric, area under the receiver operating characteristic (ROC) curve [36], to quantify the accuracy of link prediction algorithms. It is usually abbreviated as AUC value. This metric can be interpreted as the probability that a randomly chosen missing link (a link in E^P) is given a higher score than a randomly chosen nonexistent link (a link in $U \setminus E$). In the implementation, among n times of independent comparisons, if there are n' times the missing link having higher score and n'' times the missing link and nonexistent link having

the same score, we define the AUC value as [15]:

$$\text{AUC} = \frac{n' + 0.5n''}{n}.$$

If all the scores are generated from an independent and identical distribution, the AUC value should be about 0.5. Therefore, the degree to which the AUC value exceeds 0.5 indicates how much better the algorithm performs than pure chance.

IV. EXPERIMENTS

This section first introduces the basic information about studied data, and then present the experimental results as well as what we can gain from these experiments.

A. Data Description

Our experiments include 15 real directed networks drawn from disparate fields. Details are as follows and the basic structural features are presented in Table I. If a network is unconnected, we only consider its largest weakly connected component.

Biological networks. Three of them are food webs, representing the predator-pray relations, and another one is a neural network of *C.elegans*.

- FW1 [37] — A food web consists of 69 species living in Everglades Graminoids during wet season.
- FW2 [38] — A food web consists of 97 species living in Mangrove Estuary during wet season.
- FW3 [39] — A food web consists of 128 species living in Florida Bay during dry season.
- *C.elegans* [40] — A neural network of the nematode worm *C.elegans*, in which an edge joins two neurons if they are connected by either a synapse or a gap junction.

Information networks. We consider networks of documents where a directed link from i to j means the document i cites the document j , and a network of weblogs where a directed link stands for a hyperlink.

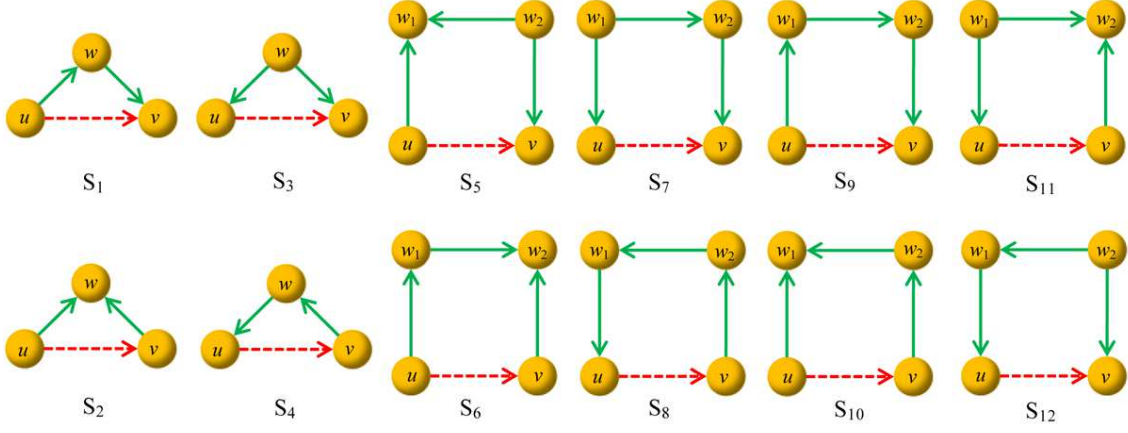


FIG. 4: Illustration of the twelve predictors corresponding to the subgraphs shown in figure 3. The red dashed arrows represent the links removed from the original subgraphs. The relations are as follows: $\{S_1, S_2, S_3\} \iff 3\text{-FFL}$, $\{S_4\} \iff 3\text{-Loop}$, $\{S_5\} \iff \text{Bi-fan}$, $\{S_6, S_7\} \iff \text{Bi-parallel}$, $\{S_8\} \iff 4\text{-Loop}$, $\{S_9, S_{10}, S_{11}, S_{12}\} \iff 4\text{-FFL}$.

- Small & Griffith and Descendants (SmaGri) [41] — Citations to Small & Griffith and Descendants.
- Kohonen [41] — Articles with topic “self-organizing maps” or references to “Kohonen T”.
- Scientometrics (SciMet) [41] — Articles from or citing Scientometrics.
- Political Blogs (PB) [42] — A directed network of hyperlinks between weblogs on US political blogs.

Social networks. All the following networks describe relationships between people.

- Delicious [43] — Delicious.com, previously known as del.icio.us, allows individuals to tag the bookmarks and follow other users. The studied who-follow-whom network was collected at May 2008.
- Youtube [44] — YouTube offers the greatest platform where users can share videos with others. Active users who regularly upload videos maintain a channel pages. Other users can follow those users thus forming a social network. This data was collected at January 2007.
- FriendFeed [45] — FriendFeed is an aggregator that consolidates the updates from the social media and social networking websites, social bookmarking websites, blogs and micro-blogging updates, etc. Members can manage their social networking contents with one Friend-Feed account and follow others’ updates. This dataset captures the who-follow-whom relationships.
- Epinions [46] — Epinions.com is a who-trust-whom online social network of a general consumer review site. Members of this site can decide whether to “trust” each other.

- Slashdot [47] — Slashdot.org is a technology-related news web site known for its specific user community. This site allows individuals to tag each other as friends or foes.
- Wikivote [34, 48] — Wikipedia is a free encyclopedia written collaboratively by volunteers around the world. Active users can be nominated to be administrator. A public voting begins after some users are nominated. Other users can express their positive, negative or neutral idea towards all the candidates. The most voted candidate will be promoted to admin status. This process implies a social network in which users are nodes and the action of voting from someone to another demonstrates a directed link. This data is from English Wikipedia on 2794 elections.
- Twitter [49] — Twitter is an online social networking service where users can post texts within 140 characters. It also allow users to “follow” other users whereby a user can see updates from the users he follows on his twitter page.

B. Results

Table 2 shows the predicting accuracies, measured by AUC values, of all the 12 individual predictors. In 14 out of 15 real networks, except Youtube, the predictor S_5 performs best among all the 12 predictors. The difference between the AUC values of the predictor S_5 and those of other predictors are usually remarkable, while for Youtube, the performance of S_5 is very close to the optimal one, S_{12} . The last row of Table 2 shows the average AUC values, which again emphasizes the advantage of S_5 . Roughly speaking, the very simple rule — a link

TABLE I: The basic structural features of the studied 15 real networks. $|V|$ and $|E|$ are the number of nodes and links, k_{max}^{in} and k_{max}^{out} are the maximum of in-degree and out-degree of all nodes, and $\langle k \rangle$ is the average degree of all nodes (average in-degree equals average out-degree). $\langle d \rangle$ and C are the 90-percentile effective diameter [50] and the clustering coefficient for directed networks [51].

Networks	$ V $	$ E $	k_{max}^{in}	k_{max}^{out}	$\langle k \rangle$	$\langle d \rangle$	C	References
FW1	69	916	63	44	13.3	2.84	0.552	[37]
FW2	97	1492	90	46	15.4	2.86	0.468	[38]
FW3	128	2137	110	63	16.7	2.90	0.335	[39]
C.elegans	297	2345	134	39	7.9	3.85	0.292	[40]
SmaGri	1024	4919	89	232	4.8	4.61	0.302	[41]
Kohonen	3704	12683	51	735	3.4	5.64	0.252	[41]
SciMet	2678	10381	121	104	3.9	6.40	0.174	[41]
PB	1222	19021	337	256	15.6	4.08	0.320	[42]
Delecious	571686	1668233	2767	11168	2.9	8.65	0.202	[43]
Youtube	1134890	4942035	25519	28644	4.4	7.17	0.081	[44]
FriendFeed	512889	19810241	31045	96659	38.6	4.92	0.215	[45]
Epinions	75877	508836	3035	1801	6.7	6.45	0.138	[46]
Slashdot	77360	828161	2539	2507	10.7	5.62	0.056	[47]
Wikivote	7066	103663	457	893	14.7	4.77	0.142	[34, 48]
Twitter	465018	834799	199	500	1.8	5.95	0.015	[49]

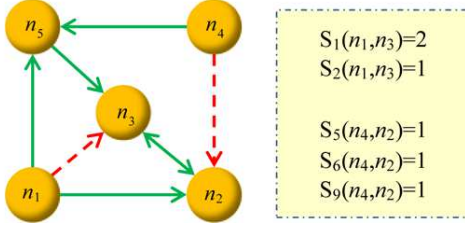


FIG. 5: Illustration of the scores of links according to our method. The red dash arrows are probe links. If we adopt the predictor S_1 , the scores for $n_1 \rightarrow n_3$ and $n_4 \rightarrow n_2$ are $S_1(n_1 \rightarrow n_3) = 2$ ($n_1 \rightarrow n_5 \rightarrow n_3$ and $n_1 \rightarrow n_2 \rightarrow n_3$) and $S_1(n_4 \rightarrow n_2) = 0$, respectively. More examples are as follows: $S_2(n_1 \rightarrow n_3) \blacktriangleright \{n_1 \rightarrow n_2 \leftarrow n_3\}$; $S_5(n_4 \rightarrow n_2) \blacktriangleright \{n_4 \rightarrow n_5 \leftarrow n_1 \rightarrow n_2\}$; $S_6(n_4 \rightarrow n_2) \blacktriangleright \{n_4 \rightarrow n_5 \rightarrow n_3 \leftarrow n_2\}$; $S_9(n_4 \rightarrow n_2) \blacktriangleright \{n_4 \rightarrow n_5 \rightarrow n_1 \rightarrow n_2\}$;

generates more Bi-fan subgraphs has higher probability to exist — is nearly 90% right.

Table 3 compares the predicting accuracies of some hybrid predictors. We explain again that the predictor $S_1 + S_2 + S_3$ means that the score of a non-observed link is defined as the number of created S_1 , S_2 and S_3 resulted from the addition of this link. In fact, the six predictors in Table 3 correspond to the six minimal loop-embedded subgraphs in figure 3. Therefore, Table 3 directly compares the six candidate subgraphs: which one is the most preferred structure so that a link generating more such structures has higher probability to exist. Again, Bi-fan wins.

Looking at the results presented in Table 2 and Table 3, a significant advantage of the Bi-fan structure is the high robustness, that is to say, even when the predictor S_5 is not the best in some cases, its performance is very close to the optimal one. In contrast, for any other predictor,

no matter an individual predictor or a hybrid one, it is very sensitive to the network structure, and occasionally give very bad predictions.

V. CONCLUSION AND DISCUSSION

This article studied the underlying mechanism on the link formation of directed networks. We presented a hypothesis named potential theory, which claims that a link that can generate more potential-definable subgraphs is of higher probability to appear. This mechanism cannot be solely used to infer network structure since there are too many potential-definable subgraphs (e.g., directed paths of any lengths are potential definable). Therefore, we also take into account two well-known local mechanisms: clustering and homophily. By combining these three mechanisms, we deduced that Bi-fan is the most preferred subgraph in directed networks. Via comparison of the link prediction accuracies of 12 individual predictors as well as six minimal loop-embedded subgraphs, Bi-fan performs best: not only for its remarkably higher AUC value than others, but also for its robustness, namely for disparate testing networks, its performance is either the best or very close to the best.

The local driven mechanisms underlying directed network formation are less understood compared with those for undirected networks. This kind of study is thus of theoretical significance, and our work provided insights into the microscopic architecture of directed networks. Although the potential theory is more complicated than the clustering and homophily mechanisms as well as the balance theory, its meaning is easy to be captured, that is, the potential-definable property implies a local hierarchy and the potential value of a node indicates its level in the hierarchical structure. For example, the di-

TABLE II: AUC values of the 12 predictors shown in figure 4. The best performance for each network is emphasized in bold. Each number is obtained by averaging over 50 implementations with independently random partitions of training set and testing set.

Datasets	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈	S ₉	S ₁₀	S ₁₁	S ₁₂
FW1	0.7400	0.4634	0.6156	0.4903	0.9066	0.6147	0.7811	0.4172	0.7848	0.4254	0.3236	0.5697
FW2	0.7629	0.5507	0.6367	0.4809	0.8964	0.6965	0.7838	0.4972	0.6822	0.4255	0.3818	0.5456
FW3	0.7333	0.5364	0.5675	0.3997	0.9105	0.7282	0.7757	0.4303	0.6683	0.3517	0.3210	0.4532
C.elegans	0.7886	0.7127	0.7569	0.5671	0.8679	0.7686	0.7991	0.5755	0.7990	0.6528	0.6667	0.7591
SmaGri	0.7074	0.6517	0.6905	0.4922	0.8852	0.7108	0.7476	0.4851	0.6677	0.6242	0.5982	0.5761
Kohonen	0.6693	0.6124	0.6642	0.4991	0.8605	0.6333	0.7335	0.4985	0.6148	0.5614	0.5778	0.5946
SciMet	0.6462	0.6192	0.6371	0.4980	0.8371	0.6672	0.7045	0.4968	0.5977	0.5794	0.5753	0.5895
PB	0.9025	0.8181	0.8243	0.6948	0.9595	0.8659	0.8679	0.7518	0.9479	0.8349	0.7616	0.8584
Delecious	0.7298	0.7077	0.7192	0.6577	0.7839	0.7141	0.7344	0.6739	0.7378	0.7081	0.7046	0.7273
Youtube	0.7518	0.7453	0.7522	0.7456	0.8517	0.8422	0.8576	0.8442	0.8505	0.8430	0.8507	0.8624
FriendFeed	0.8801	0.7503	0.7382	0.5895	0.9766	0.7863	0.8100	0.7150	0.9690	0.8324	0.7318	0.8027
Epinions	0.8273	0.8326	0.8081	0.7460	0.9101	0.8969	0.8843	0.8584	0.8995	0.8956	0.8804	0.8831
Slashdot	0.7164	0.7133	0.7124	0.7072	0.9035	0.8984	0.8982	0.8925	0.9009	0.8982	0.8926	0.8985
Wikivote	0.9073	0.7448	0.7470	0.5962	0.9699	0.7679	0.7451	0.6209	0.9583	0.7562	0.6096	0.7468
Twitter	0.5211	0.5033	0.5182	0.5002	0.7498	0.5033	0.5441	0.5002	0.5549	0.5037	0.5039	0.5423
Average	0.7523	0.6641	0.6925	0.5776	0.8846	0.7396	0.7778	0.6172	0.7756	0.6595	0.6253	0.6940

TABLE III: AUC values of the six subgraphs shown in figure 3. The best performance for each network is emphasized in bold. Each number is obtained by averaging over 50 implementations with independently random partitions of training set and testing set.

Datasets	S ₁ + S ₂ + S ₃	S ₄	S ₅	S ₆ + S ₇	S ₈	S ₉ + S ₁₀ + S ₁₁ + S ₁₂
FW1	0.6953	0.4903	0.9066	0.8462	0.4172	0.4653
FW2	0.7241	0.4809	0.8964	0.8490	0.4972	0.4674
FW3	0.6649	0.3997	0.9105	0.8586	0.4303	0.3283
C.elegans	0.8666	0.5671	0.8679	0.8403	0.5755	0.7736
SmaGri	0.8400	0.4922	0.8852	0.8154	0.4851	0.7291
Kohonen	0.8091	0.4991	0.8605	0.7779	0.4985	0.7039
SciMet	0.7874	0.4980	0.8371	0.7872	0.4968	0.7187
PB	0.9275	0.6948	0.9595	0.9029	0.7518	0.9122
Delecious	0.7621	0.6577	0.7839	0.7743	0.6739	0.7893
Youtube	0.7526	0.7456	0.8517	0.8593	0.8442	0.8625
FriendFeed	0.7937	0.5895	0.9766	0.9151	0.7150	0.9240
Epinions	0.8682	0.7460	0.9101	0.9131	0.8584	0.9174
Slashdot	0.7422	0.7072	0.9035	0.9048	0.8925	0.9083
Wikivote	0.9330	0.5962	0.9699	0.8607	0.6209	0.9288
Twitter	0.5281	0.5002	0.7498	0.5475	0.5002	0.5733
Average	0.7797	0.5776	0.8846	0.8302	0.6172	0.7335

rected loops are not hierarchy-embedded and the directed path is of a strictly hierarchical organization, the former is not potential-definable and the later is potential-definable. The hierarchical organization is a well-known macroscopic feature for many undirected [52, 53] and directed [54] networks, our work indicates that for directed networks, nodes tend to locally self-organized in a hierarchical manner. We guess this kind of microscopic hierarchical organization will contribute to the macroscopic hierarchical structure. In the near future, we will study more data sets in a more detailed way to check whether the potential theory and our hypothesis on hierarchical organization are valid, and to see the applicable range (to which networks it works and to what extent it can explain the network formation) of the potential theory.

Lastly, we would like to say again that the link pre-

diction problem is a very fundamental problem for both information filtering and network analysis [15, 55], and it could find countless applications. In this work, we applied the link prediction approach to evaluate driven mechanisms of network formation, at the same time, our method can be directly applied in predicting missing links and recommending friendships for large-scale directed networks, since the accuracy of our method is much higher than the common-neighbor-based methods as indicated by the performance of predictors S_1 , S_2 , S_3 and S_4 .

Acknowledgments. We acknowledge An Zeng, Changsong Zhou and Xiao-Ke Xu for their helpful discussions and irradiative ideas. This work is partially supported by the National Natural Science Foundation of China under Grant No. 11075031, and the Swiss National

-
- [1] M. E. J. Newman. *Networks: An Introduction*, (Oxford University Press, New York, 2010).
- [2] A.-L. Barabási. Scale-free Networks: A decade and beyond. *Science*, 325 (2009) 412.
- [3] A.-L. Barabási, R. Albert. Emergence of scaling in random networks. *Science*, 286 (1999) 509.
- [4] T. Zhou, M. Medo, G. Cimini, Z.-K. Zhang, Y.-C. Zhang. Emergence of scale-free leadership structure in social recommender systems. *PLoS ONE*, 6 (2011) e2064.
- [5] J. I. Perotti, O. V. Billoni, F. A. Tamarit, D. R. Chialvo, S. A. Cannas. Emergent self-organized complex network topology out of stability constraints. *Physical Review Letters*, 103 (2009) 108701.
- [6] M. McPherson, L. Smith-Lovin, J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27 (2001) 415.
- [7] G. Szabó, M. Alava, J. Kertész. Clustering in complex networks. *Lecture Notes in Physics*, 650 (2004) 139.
- [8] S. A. Marvel, S. H. Strogatz, J. M. Kleinberg. The energy landscape of social balance. *Physical Review Letters*, 103 (2009) 198701.
- [9] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44-54, 2006.
- [10] G. Palla, A.-L. Barabási, T. Vicsek. Quantifying social group evolution. *Nature*, 446 (2007) 664.
- [11] J. M. Kumpula, J.-P. Onnela, J. Saramäki, K. Kaski, J. Kertész. Emergence of communities in weighted networks. *Physical Review Letters*, 99 (2007) 228701.
- [12] P. Holme, B. J. Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65 (2002) 026107.
- [13] D. Yin, L. Hong, X. Xiong, B. D. Davison. Link formation analysis in microblog. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1235-1236, 2011.
- [14] D. Garlaschelli, M. I. Loffredo. Patterns of link reciprocity in directed network. *Physical Review Letters*, 93 (2004) 268701.
- [15] L. Lü, and T. Zhou. Link Prediction in Complex Networks: A Survey. *Physica A*, 390 (2011) 1150.
- [16] D. J. Watts, S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (1998) 440-442.
- [17] G. Kossinets, and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311 (2006) 88.
- [18] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462-470, 2008.
- [19] A.-X. Cui, Y. Fu, M.-S. Shang, D.-B. Chen, T. Zhou. Emergence of local structures in complex networks: Common neighborhood drives the network evolution. *Acta Physica Sinica*, 60 (2011) 038901.
- [20] J. Leskovec, E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915-924, 2008.
- [21] S. Currarini, M. O. Jackson, P. Pin. Identifying the roles of race-based choice and chance in high school friendship network formation. *Proc. Natl. Acad. Sci. U.S.A.*, 107 (2010) 4857.
- [22] K. Lewis, M. Gonzalez, J. Kaufman. Social selection and peer influence in an online social network. *Proc. Natl. Acad. Sci. U.S.A.*, 109 (2012) 68.
- [23] X.-Q. Cheng, F.-X. Ren, S. Zhou, M.-B. Hu. Triangular clustering in document networks. *New Journal of Physics*, 11 (2008) 033019.
- [24] M. J. Brzoowski, D. M. Romero. Who should I follow? Recommending people in directed social networks. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 458-461, 2011.
- [25] T. Opsahl, B. Hogan. Modeling the evolution of continuously-observed networks: Communication in a Facebook-like community. *e-print arXiv: 1010.2141*, 2010. <http://arxiv.org/abs/1010.2141>.
- [26] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, B. Bhattacharjee. Growth of the flickr social network. In *Proceedings of the first workshop on Online social networks*, pages 25-30, 2008.
- [27] V. Gómez, A. Kaltenbrunner, V. López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*, pages 645-654, 2008.
- [28] S. L. Pimm. *Food Webs*, University of Chicago Press, Chicago, 2002.
- [29] L. Getoor, C. P. Diehl. Link mining: A survey. *ACM SIGKDD Explor. Newsl.*, 7 (2005) 3.
- [30] D. Liben-Nowell, J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58 (2007) 1019.
- [31] T. Zhou, L. Lü, Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71 (2009) 623.
- [32] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on Web*, (to be published), 2012.
- [33] D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabási. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages: 1100-1108, August 2011.
- [34] J. Leskovec, D. Huttenlocher, J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. In *Proceedings of the 19th international conference on World wide web*, pages 641-650, 2010.
- [35] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298 (2002) 824.
- [36] J. A. Hanely, B. J. McNeil. The meaning and user of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 (1982) 29.
- [37] R. E. Ulanowicz, J. J. Heymans, M. S. Egnor-

- tovich. Network Analysis of Trophic Dynamics in South Florida Ecosystems, FY 99: The Graminoid Ecosystem. Technical report, TS-191-99, 2000. <http://www.cbl.umces.edu/~atlss/swgras701.html>.
- [38] D. Baird, J. Luczkovich, R. R. Christian. Assessment of spatial and temporal variability in ecosystem attributes of the St Marks National Wildlife Refuge, Apalachee Bay, Florida. *Estuarine, Coastal and Shelf Science*, 47 (1998) 329.
- [39] R. E. Ulanowicz, C. Bondavalli, M. S. Egnotovich. Network Analysis of Trophic Dynamics in South Florida Ecosystem, FY 97: The Florida Bay Ecosystem. Technical report, CBL 98-123, 1998. <http://www.cbl.umces.edu/~atlss/FBay701.html>.
- [40] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode *C. elegans*. *Phil. Trans. Royal Soc. London. Series B, Biol. Scien.*, 314 (1986) 1.
- [41] <http://vlado.fmf.uni-lj.si/pub/networks/data/>.
- [42] L. A. Adamic, and N. Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36-43, 2005.
- [43] L. Lü, Y.-C. Zhang, C. H. Yeung, T. Zhou. Leaders in social networks, the Delicious case. *PLoS ONE*, 6 (2011) e21202.
- [44] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29-42, 2007.
- [45] F. Celli, F. M. L. Di Lascio, M. Magnani, B. Pacelli, L. Rossi. Social network data and practices: the case of friendfeed. In *International Conference on Social Computing, Behavioral Modeling, and Prediction*, pages 346-353, 2010.
- [46] M. Richardson, R. Agrawal, P. Domingos. Trust Management for the Semantic Web. In *Proceedings of the 2nd International Semantic Web Conference*, 2003.
- [47] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics*, 6 (2009) 29.
- [48] J. Leskovec, D. Huttenlocher, J. Kleinberg. Signed Networks in Social Media. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1361-1370, 2010.
- [49] M. De Choudhury, Y. -R. Lin, H. Sundaram, K. S. Candan, L. Xie, A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, May 2010.
- [50] C. R. Palmer, P. B. Gibbons, C. Faloutsos. Anf: A fast and scalable tool for data mining in massive graphs. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81-90, 2002.
- [51] G. Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76 (2007) 026107.
- [52] A. Clauset, C. Moore, M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453 (2008) 98.
- [53] A. Lancichinetti, S. Fortunato, J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11 (2009) 033015.
- [54] H. Yu, M. Gerstein. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103 (2006) 14724.
- [55] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou. Recommender Systems. *Physics Reports* (to be published), 2012.